DOCUMENT RESUME

ED 192 446 EA 012 939

AUTHOR Pratt, Harold; And Others
TITLE The Effects of a Concerns-Based Implementation Plan on the Achievement of Elementary Science Students.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Apr 80
NOTE 23p.: Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 7-11, 1980). Some figures may not reproduce clearly due to small print of original document.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Curriculum Development; Elementary Education; *Elementary School Science; *Instructional Innovation; Reading Comprehension; Test Construction; *Test Results; *Time Factors (Learning)

ABSTRACT

Recently the framework of the Concerns-Based Adoption Model was used to guide the implementation of a revised elementary science program in grades 3-6 in the Jefferson County (Colorado) Public Schools. This paper explores how student achievement on a locally developed, criterion-referenced test is related to the length of time the revised science program has been in use. Three groups of students drawn from schools in three phases of the implementation were identified based on reading scores. The data indicate a definite trend toward increased student achievement in those classrooms where the innovation had been implemented for a longer time period. (Author/MLF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED192446

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE EFFECTS OF A CONCERNS-BASED IMPLEMENTATION PLAN ON THE ACHIEVEMENT OF ELEMENTARY SCIENCE STUDENTS

Harold Pratt

Susie Winters

Jefferson County Public Schools
Lakewood, Colorado

Archie George

Research and Development Center for Teacher Education
The University of Texas at Austin

Paper presented at the annual meeting of the American Educational Research Association, Boston, April, 1980

EA 012 939

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. George

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE EFFECTS OF A CONCERNS-BASED IMPLEMENTATION PLAN ON THE ACHIEVEMENT OF ELEMENTARY SCIENCE STUDENTS[1]

Harold Pratt

Susie Winters

Archie George

Introduction

The ultimate purpose for the implementation of a new curriculum is the improvement of the educational opportunity of the students involved. This usually means changing a program presented to students or improving student achievement in a given content or subject area. Recently the framework of the Concerns-Based Adoption Model (Pratt, Melle, Metzdorf & Loucks, 1980) was used to guide the implementation of a revised elementary science program in grades 3-6 in the Jefferson County, Colorado Public Schools. One of the previous findings of research related to the model is that the level at which an innovation is used improves or increases over time measured in terms of years (Hall & Rutherford, 1976). Since one of the major assumptions of CBAM is that the implementation of an innovation should be considered a process, not an event, the Level of Use (LoU) of the revised elementary science program was

---

[1]The research described herein was conducted under contract with the National Institute of Education. The opinions expressed are those of the authors and do not necessarily reflect the position or policy of the National Institute of Education, and no endorsement by the National Institute of Education should be inferred.

monitored for more than two years before, during and after the year-long in-service plan. The results of this study (Loucks & Melle, 1980) show that the Level of Use of the revised program did increase over the period of time of the study. This paper explores how student achievement on a locally developed, criterion-referenced test is related to the length of time the revised science program has been in use.

As previously stated, CBAM assumes change to be a process. In order to provide more time for the process to occur, three one-day sessions of inservice were scheduled approximately three months apart, allowing teachers an opportunity to use the new program between sessions. (See Pratt, Melle, Metzdorf & Loucks, 1980, for more details of the inservice plan.) Because of the large number (approximately 700) of elementary teachers involved, the district was divided into three equal groups of schools, or phases. Phase II and Phase III began their inservice programs about four months and twelve months, respectively, after Phase I. This "phasing in" over time of the revised program provided an opportunity to measure the difference between student achievements in classrooms whose teachers had implemented the new program the earliest, and classrooms whose teachers started using the program at a later time. The time difference in implementation was as long as one year. This allowed direct measurement of the relationship between student achievement and length of time that a teacher had been using the new material.

Test Development Procedures

The Jefferson County elementary science program consists almost entirely of "hands-on" type units from either Elementary Science Study (ESS) or Science Curriculum Improvement Studies (SCIS). Because of its nature, the program

developers considered it inappropriate to measure student achievement within such a curriculum by a multiple choice test. It was thought perhaps a commercially available standardized test had been developed which correlated inquiry skills or processes that could be directly observed during a "hands-on" lesson. A search of the literature in 1974 by the Bureau of Educational Field Services at the University of Colorado failed to produce such a test (Riley, 1975).

The science department therefore tackled the task of creating a sixth grade test that would be an adequate measure of student achievement in a very active, "hands-on" inquiry curriculum. To do so, two essentially parallel versions of the same test were created. The first was a "hands-on," one-to-one test in which the student manipulated materials and equipment to display a mastery of specific science concepts or understandings. Such a test would be highly unfeasible to administer in a large school district because of cost involved. Therefore, a second version of the test was created in a multiple-choice format. If a strong correlation could be found between these two tests then it could be assumed that both tests were measuring the same construct, and the easier to administer multiple choice test could be used. If the correlation did not exist, then a small random sample of students could be administered this one-to-one test, and district results extrapolated from their results.

The first step in the test development was to identify those concepts representative of the elementary program in grades 3-6. Science inservice leaders (classroom teachers who had been trained in conducting science inservice classes) and those who had helped write and pilot-test the new curriculum were asked to identify specific concepts from each unit they thought were important for students to know. Once this list of "objectives" was completed, this same group

of people wrote "hands-on," one-to-one items for each of the specific concepts. At the same time they wrote corresponding multiple choice items matching the "hands-on" items. After the items were reviewed a number of times the two versions of the test were prepared for pilot-testing.

Pilot Test Procedures and Results

Three pilot-test schools were chosen so as to provide students with a wide range of ability, socio-economic population and previous experience with the science program. Fifty-nine students were randomly selected from nine sixth grade classrooms. Three testers were trained to administer the "hands-on" test and their reliability informally established. Due to the length of the "hands-on" test each student received three-fourths of the "hands-on" items. The multiple choice was administered by one of the testers by being read aloud during a 75 minute testing session. Half of the sample received the multiple choice before the "hands-on" test and the other half after. Twenty-four items from the National Assessment of Educational Progress (NAEP) Science Tests for nine and thirteen year olds were embedded in the multiple choice test. These items were selected because they had been carefully screened and evaluated for use in the national instrument, and because their content coverage was very closely related to the district program. Ultimately, student scores on these items could be used for a limited comparison between local achievement results and those at the national level. In addition, CTBS reading comprehension scores were available on each student from tests that had been administered a few months earlier.

Norris Harms (1978) of the Bureau of Educational Field Services of the University of Colorado executed an extensive item and test analysis. Several scales were defined in the areas of physical science, environmental science, astronomy and earth science, and health science, as well as the NAEP multiple

choice items and the locally developed multiple choice items. Figure 1 shows the number of items on each scale, an estimate of each scale's reliability, mean and standard deviation and the number of students tested. Because each student in the sample responded to only approximately three-fourths of the one-to-one test items, the reliability of the entire one-to-one scale was estimated by using the Spearman-Brown formula. This calculation is displayed in Figure 2.

Figure 1

SCALE	NUMBER OF ITEMS	RELI-ABILITY	MEAN	STANDARD DEVIATION	NUMBER STUDENTS TESTED
Health 1-1	25	.72	13.9	3.72	40
Physical Science 1-1	16	.58	8.6	2.58	42
Environment 1-1	18	.52	9.7	2.64	42
Astronomy - Earth	14	.25	6.6	1.76	42
Total 1-1	73	.80*			
Jeffco Multiple Choice	71	.78	42.0	7.90	59
NAEP Multiple Choice	24	.80	16.3	4.37	59
All Multiple Choice	95	.86	58.4	11.32	59
CTBS			53.2	70.1	

*Estimated using the Spearman-Brown formula (See Figure 2).

Figure 2

RELIABILITY ESTIMATES - ONE-TO-ONE SCALE

GROUP I - (Health and Physical Science and the Environment) r_{tt}=.79

GROUP II - (Health and the Environment and Earth/Astronomy) r_{tt}=.75

GROUP III - (Physical Science and the Environment and Earth/Astronomy) r_{tt}=.63

GROUP IV - (Health and Physical Science and Earth/Astronomy) r_{tt}=.82

Mean r_{tt}=.75

Estimate of r_{tt} for whole test (r_{tt}=.75 adjusted for a length of 4/3 by Spearman Brown formula).

$$n = \frac{nr_{tt}}{1 + (n-1)r_{tt}} = \frac{4/3\ (.75)}{1 + (1/3)\ (.75)} = \frac{1}{1.25} = .80$$

Our best estimate of r_{tt} is .80. This estimate should be fairly accurate, with little error due to sampling.

A correlation matrix between the individual scales is provided in Figure 3. It should be noted that the multiple choice scores correlate somewhat higher with the CTBS reading scores than do the "hands-on," one-to-one scores. Thus it appears that performance on the one-to-one scale is somewhat less dependent on reading than are the paper and pencil multiple choice scores. It is notable that the district one-to-one test correlates higher with the district multiple choice items than with the NAEP multiple choice items. This indicates that the Jeffco multiple choice scale comes closer to measuring that domain identified by the one-to-one scale and, therefore, the district curriculum as presented to students, than does the NAEP scale.

The .66 correlation between the district multiple choice items and the district one-to-one items indicates an acceptable degree of agreement between the two scales. Thus, to the extent that the one-to-one test measures the district objectives in curriculum program, the district multiple choice test also measures these objectives. We therefore decided to measure student achievement using the multiple choice test, not the one-to-one test.

Figure 3

SELECTED CORRELATION MATRIX - COGNITIVE AND ONE-TO-ONE SCALES

	Jeffco 1-1 Average**	Jeffco M.C.	NAEP M.C.	All M.C.	CTBS Reading
Jeffco 1 to 1	1.0	.66	.48	.66*	.31
Jeffco M.C.	.66	1.0	.69	.96	.46
NAEP M.C.	.48	.69	1.0	.87*	.38
All M.C.	.66	.96*	.87*	1.0	--
CTBS Reading	.31	.46	.38	--	1.0

*Note that the "Jeffco Multiple Choice" and NAEP Multiple Choice scores are combined to produce the "all multiple choice" scores. The correlations of part to whole scores include some auto-correlation and are high.

**Average computed via Fisher Z.

Field Test

The development of the field test version of the multiple choice test during the second year (1978-79) concentrated on the reduction and refinement of the multiple choice items. The field test version could also be used to compare student achievement among the three phases of implementation. The three phases represent the schedule for which teachers attended the inservice

for, and subsequently began teaching, the revised elementary science program.

Phases	Dates of Inservice	Months of Experience with Revised Program at Time of Test Administration
I	Jan 1977-Sept 1977	29
II	Mar 1977-Sept 1977	25
III	Jan 1978-Sept 1978	17

The students from the phases were matched by selecting a sample from each one so that the distribution of CTBS reading comprehension scores was the same as the distribution of scores in the total district. Figure 4 displays the scores for CTBS reading comprehension of the sample populations. Each of the twelve schools administered their own test with students reading the items themselves with no time limit.

Figure 4

CTBS - 1978-79

Grade 6

Reading Comprehension (ESS)

Sample Population

	N	X	S.D.
TOTAL	174	558.1	80.3
Phase 1	59	558.9	78.3
Phase 2	61	556.6	80.8
Phase 3	54	559.0	81.9

Kuder-Richardson formula 20 estimates of internal consistency are shown in Figure 5.

Figure 5

Scale	Number of Items	Reliabilities
Health	30	.61
Physical	25	.60
Environmental	32	.76
Earth/Astronomy	13	.46
Total	100	.86

A step-down, linerar-models approach was used to analyze the data on all scales. The models used can be thought of as an expanded analysis of covariance, the expansion being tests of assumptions made in conventional analysis of covariance.

The first model used contains nine parameters. Essentially, data points from each phase were used to estimate three parameters independently. Within each phase, the model had the form:

$$Y = a_0U + a_1X + a_2X^2 + E,$$

where U is the unit vector,

X contains the CTBS reading score,

X^2 contains CTBS reading score squared,

and Y contains the dependent variable.

This model is familiar to many readers as it is used frequently to fit data with a curved line.

The first "restriction" or step-down placed on the model was to require that the estimate of the parameter on the squared CTBS score, a_2, be the same for all three groups. This resulted in a model with seven parameters, a separate a_0 and a_1 for each group and a common a_2.

The second restriction tested the hypothesis that this common parameter,

estimated using data from all three groups, was actually zero. That is, is it necessary to use curved lines, or do three straight lines fit the data as well? This restriction resulted in a model with six parameters, an independent slope, a_1, and intercept, a_0, for each group.

The third restriction tested the hypothesis that the slope of the line, estimated by a_1, was the same in all three groups. That is, is it reasonable to assume that the three lines are parallel? This restriction resulted in a model with four parameters, three separate intercepts, one for each group, and a common slope.

The fourth, and final, restriction tested the hypothesis that the intercepts were the same. That is, whether it was reasonable to assume that there was no difference between the groups. This resulted in a two parameter model, simply a slope and intercept for the entire sample.

When each restriction was imposed, an F-test was used to evaluate the increase in error when the simpler model was used to fit the data. The step-down process was not followed beyond the point the first significant ($p<.05$) F-ratio was found (Ward & Jennings, 1973). In three of the five cases (Total Score, Physical and Earth/Astronomy), the seven parameter model was significantly "better than" the six parameter model. In the other two cases (Health and Environmental), the six parameter model was determined to be superior to the four parameter model. Thus, significant differences between the groups were found in all five analyses.

Each of the analyses was replicated using the CTBS Math score as the covariate, with nearly identical results. Even though there were some differences in the value of the tests at each step, the implications of the analyses were the same, so only the analyses using the reading scores are detailed here. (See Figures 11 and 12 to compare Reading and Math as covariates.)

The analyses indicate that more experience with the science program led to increased achievement for high ability students. High ability students in Phase I schools did better than high ability students in Phase II schools, who in turn out performed high ability students in Phase III schools. For medium and low ability students, the achievement scores were essentially the same across the three phases. This interaction between ability and performance is apparent in the analysis of the Total Score (Figure 6), the Physical Strand (Figure 7) and Health Strand (Figure 8).

Figure 6

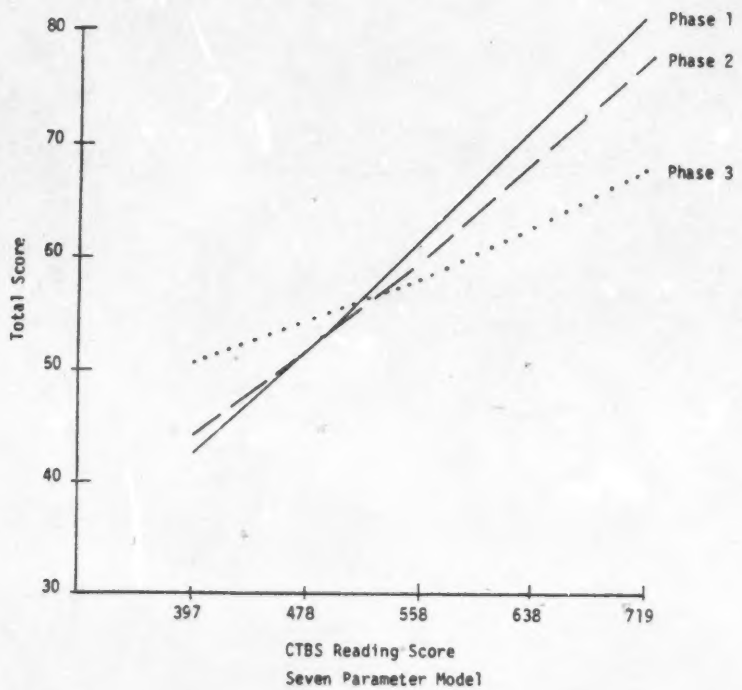


Figure 7

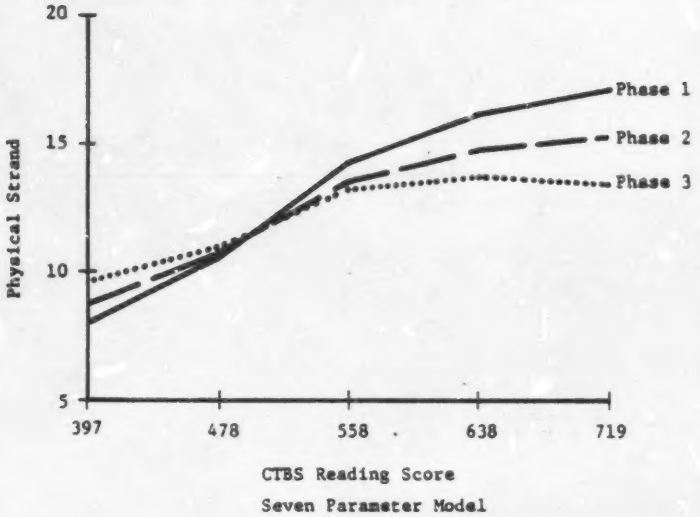
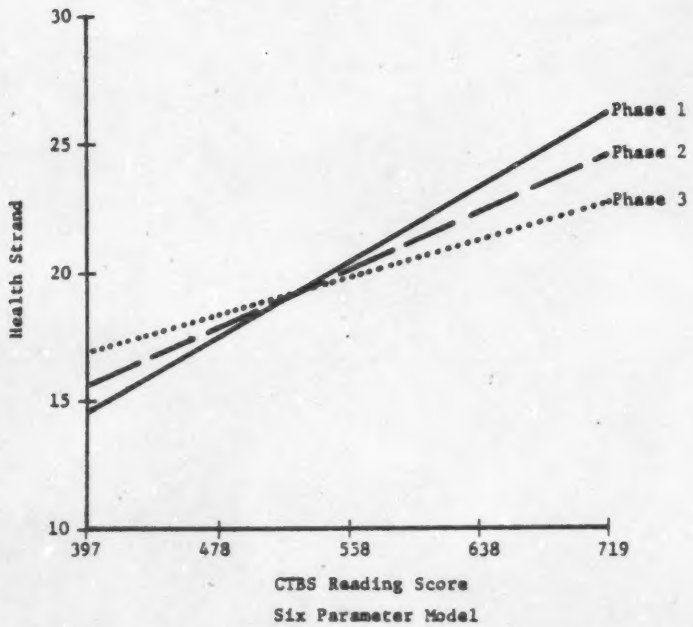


Figure 8



Somewhat different results were obtained on the other two scales. On the Environmental Strand (Figure 9), Phases I and II were essentially identical. These two groups were very similar on the Earth/Astronomy Strand (Figure 10) items also, even though Phase I students very slightly out scored Phase II at all levels of ability. In each of these subtests, a strong interaction between reading ability, performance on the science achievement test, and the phase of science implementation is apparent. In every case, Phase I and III maintain the same relationship, high ability students in Phase I do better than high ability students in Phase III. But low ability Phase III students do as well as or better than low ability Phase I students.

Figure 9

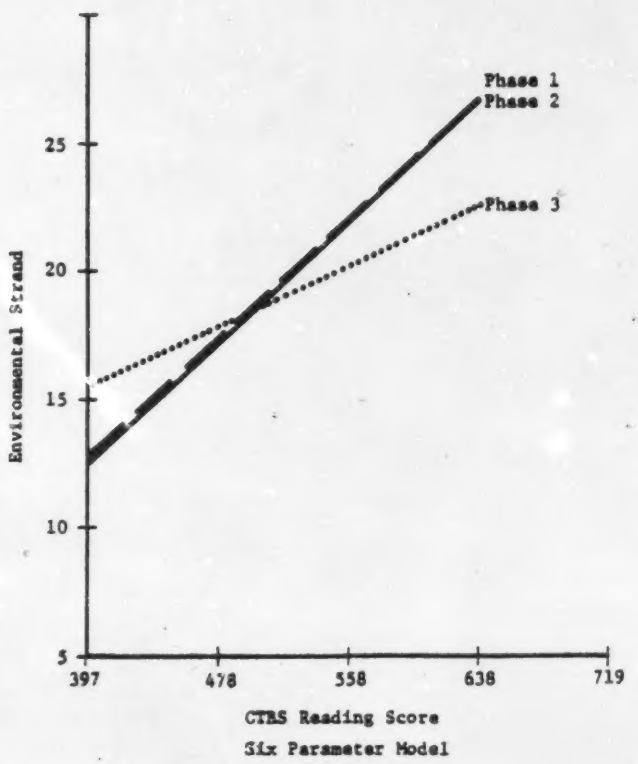
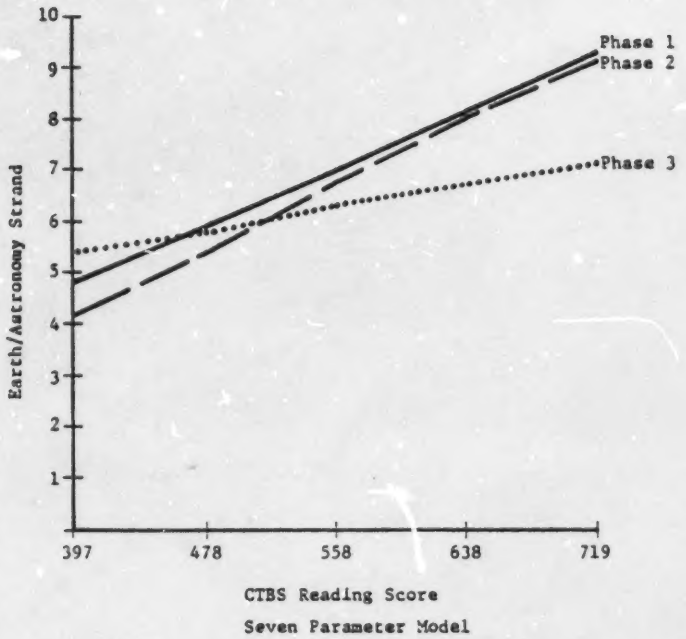


Figure 10



A procedure developed by Johnson and Neyman (1936) and programmed by Borich (1971) was used to test for "regions of significance" in the analyses of the science achievement data. The CTBS reading and math scores were used in separate analyses, and phases were compared on a pair-wise basis -- 1 and 2, 2 and 3, and 1 and 3. The Johnson-Neyman test provides cut-off scores above and below which the groups have different expected values, if the regression lines cross. Figures 11 and 12 show the results of the tests for both reading and math scores, and for each of the content areas in the science curriculum. Figures 13 and 14 show the graphical representation of the results of the test.

Comparisons of Phases II and III revealed several significant differences, especially using reading as the covariate. Phase II students of moderate to high ability did better on the Environment, Earth/Astronomy, and Total Scores than students in Phase III of equal ability. On several scales, it appears

Figure 11

	1 vs 2	2 vs 3	1 vs 3
Health:	ns	ns	3>1, R 442 1>3, R 620
Physical:	ns	ns	1>3, all R
Environmental:	ns	3>2, R<358 2>3, R>596	3>1, R<390 1>3, R>609
Earth/Astronomy:	ns	2>3, R>619	1>3, all R
Total:	ns	2>3, R>589	3>1, R<435 1>3, R>573

Results of Johnson-Neyman Test for Differences in Expected Values on Each of the Cognitive Scales Using CTBS Reading Score as the Predictor

Figure 12

	1 vs 2	2 vs 3	1 vs 3
Health:	ns	ns	ns
Physical:	ns	ns	1>3, all M
Environmental:	ns	2>3, M>559	3>1, M<406 1>3, M>544
Earth/Astronomy:	ns	ns	1>3, all M
Total:	ns	ns	1>3, M>528

Results of Johnson-Neyman Test for Differences in Expected Values on Each of the Cognitive Scales Using CTBS Math Score as the Predictor

Figure 13

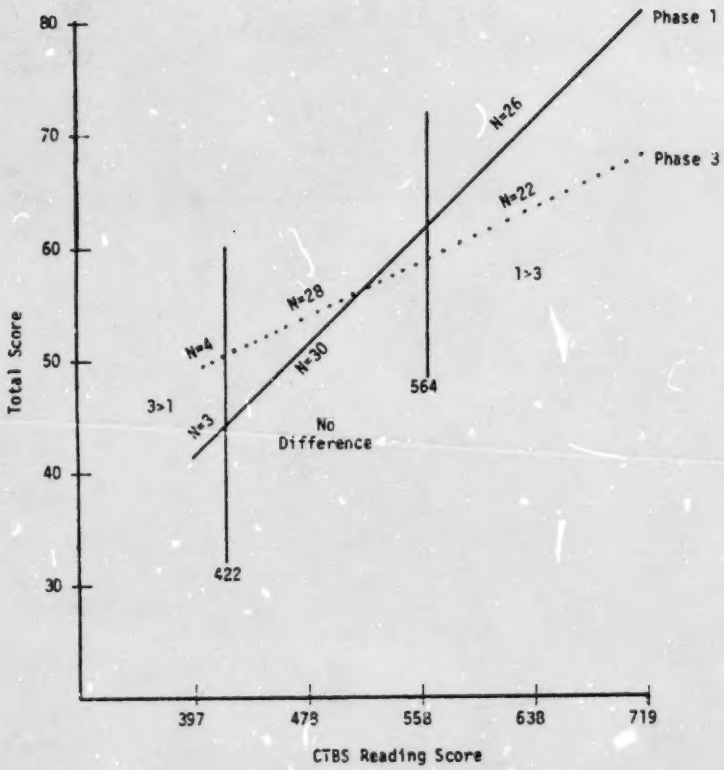
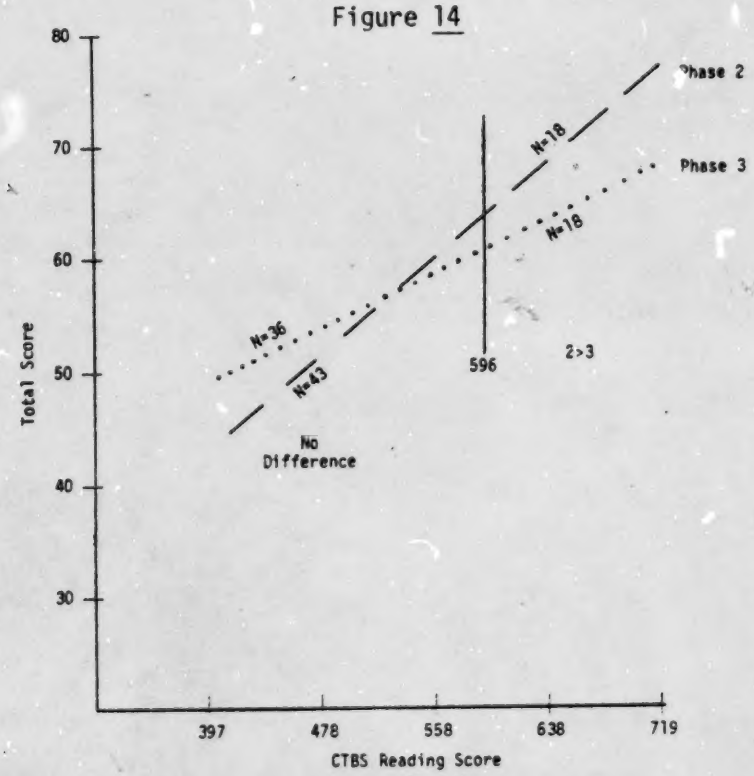


Figure 14



that students of lower ability in reading did better in Phase III than Phase II. However, this is largely illusionary -- less than 3% of the students in the sample had scores lower than 400 on the reading test. Thus, the apparent differences are based on projections of the slope of the line into this region of little data. Apparently, because the high ability students in Phase II dramatically out performed the high ability students in Phase III, and there was no difference in performance between the phases for students in the middle range, the regression lines cross. The interaction is not due to an inversion of the performance among lower ability students. The regression lines cross because of very great differences in achievement among the high ability students and the essentially equal performance of students in the low to moderate range of ability.

Very few students had CTBS scores which would be considered low in any other context. A score of 400 on the reading scale is only one standard deviation below the national average. Approximately 16% of scores are expected below this point, but only 3% are observed there. Thus, it is really not possible to say how students of low ability would perform on these tests, nor whether there would be differences between the phases in that range of ability.

Comparison of Phase I and III reveals many more differences than either of the two previous comparisons. When reading scores were used as the covariate, students of moderate to high ability in Phase I scored higher than similar students in Phase III on all five scores. Phase III students of lower ability were projected to do better than similar Phase I students on four of the five scales. Again, so few students were actually observed in the lower range that it would be misleading to infer that lower ability students in Phase III did better than those in Phase I.

Each of the four strand tests contained items which related to content scheduled to be taught in each of the four grade levels -- 3, 4, 5, or 6. Only

sixth grade students were tested. Thus, students in Phase I and II schools had very likely received science instruction under the new program in grades 4 and 5, as well as 6. Students in Phase III schools had likely received only fifth and sixth grade science under the new program.

In order to investigate whether previous years' instruction was responsible for the results described so far, a scale containing only the sixth grade items was constructed. This scale contained thirty items, nine from the Health Strand, eight from Physical Science, nine from Environmental, and four from Earth/ Astronomy. These were all of the sixth grade items on the test, including five NAEP items that were directly related to the curriculum. This scale had a mean of 16.3, standard deviation of 4.31, and reliability of .66.

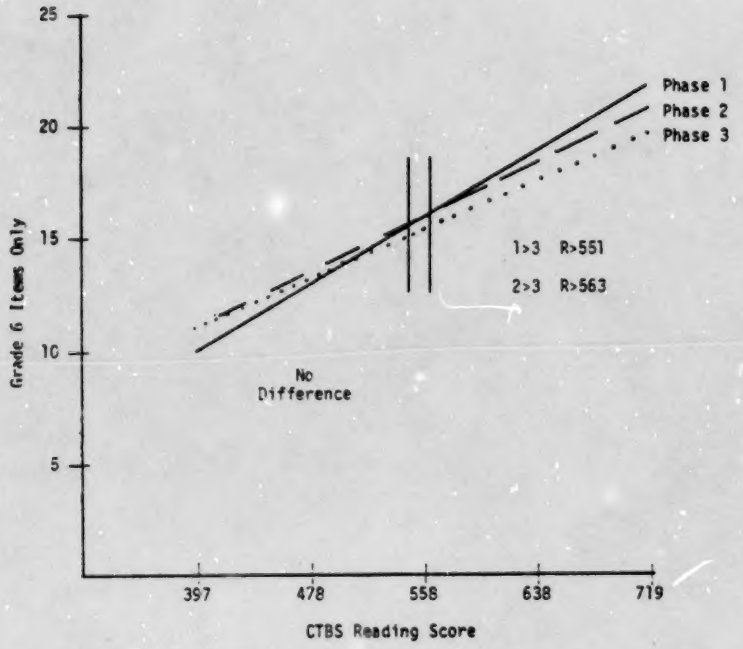
Performance on the sixth grade items only should reflect differences in the use of the program due to greater experience with the program on the part of the teacher. The sixth grade items have no prerequisite knowledge requirements from the previous grade levels that would give the students who had previously been in the program an advantage on these items.

Using reading as the covariate, the Johnson-Neyman procedure indicated that high ability Phase I students and high ability Phase II students out scored high ability Phase III students on the grade six items (Figure 15). No differences were found among moderate to low ability students nor between Phases I and II. The same results were found using CTBS math as the covariate, except no differences were found between Phases II and III.

Discussion and Implications

The analyses indicate that high ability students (those having scores above 550 on the CTBS reading test) in Phase I perform better than high ability students in Phase III on the Jeffco science test. There are no consistent differences between achievement scores for students of lower ability, nor between

Figure 15



Phases I and II or II and III.

The most straightforward interpretation of these results is that teachers with greater experience with the new science program are more effective with students of high ability, but experience does not make much difference for students of medium or low ability. This conclusion rests on the assumption that the Phase I teachers are implementing the program more fully than the Phase III teachers. Although we do not have Levels of Use data on the teachers of the students in this sample, it is true in general that the Phase I teachers had higher Level of Use data at the time of the testing than the Phase III teachers (Loucks & Melle, 1980). That is, more years of experience in use of the program

by the teacher leads to better usage which in turn leads to higher achievement by the students.

The fact that the increased teacher experience had an impact on only the high ability students may simply be caused by the fact that better students typically gain more from most instructional treatments. An alternative explanation also seems plausable to the authors -- the multiple choice test is biased in favor of those who can read well. Students with high reading ability are able to transfer understanding of the scientific process into the testing situation, which depends heavily on a high reading ability. The students with average or low reading ability are not as capable of answering the multiple choice test items as the high ability students, even though they may be learning more science and gaining a better understanding of the scientific process in the new program. This explanation leads one to expect that the one-on-one test would have shown superior achievement in Phase I at all levels of ability, not just the higher levels. It may be possible to test this hypothesis in the future.

References

Borich, G. D. Interactions among group regressions: Testing homogeneity of group regressions and plotting regions of significance. Educational and Psychological Measurement, 1971, 31, 251-253.

Hall, G. E. & Rutherford, W. L. Concerns about implementing team teaching, Educational Leadership, December 1976, 34(3), 227-33.

Harms, N. Test Analysis Report, University of Colorado Bureau of Educational Field Services, 1978.

Johnson, P. O. & Neyman, J. Tests of certain linear hypotheses and their application to some educational problems. Statistical Research Memoirs, 1936, 1, 57-93.

Loucks, S. F. & Melle, M. Implementation of a district-wide science curriculum: The effects of a three year effort. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Pratt, H. A., Melle, M., Metzdorf, J., & Loucks, S. F. The design and utilization of a concerns-based staff development program for implementing a revised science curriculum in eighty elementary schools. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Riley, J. Personal correspondence, University of Colorado Bureau of Educational Field Services, 1975.

Ward, J. H., Jr. & Jennings, E. Introduction to linear models. New York: Prentice-Hall, Inc., 1973.